Homework 5

The purpose of this homework is to practice using randomization methods to run hypothesis tests for more than two means and for correlation. Please fill in the appropriate code and write answers to all questions in the answer sections, then submit a compiled pdf with your answers through gradescope by 11:59pm on Sunday October 6th.

As always, if you need help with any of the homework assignments, please attend the TA office hours which are listed on Canvas and/or ask questions on Piazza. Also, if you have completed the homework, please help others out by answering questions on Piazza.

Part 1: Hypothesis tests for more than two means

Are movies that have particular Motion Picture Association of America (MPAA) ratings (i.e., G, PG, PG-13 or R) enjoyed more by movie critics? In the exercises below we will run hypothesis tests to examine this question using data from ~456 movies randomly selected from the Rotten Tomatoes website.

The code below loads the movie data in an object called `movies` and it also creates an object called `movies3` which only keeps movies with ratings of G, PG, PG-13 and R. For all the exercises below, **only use the data in the movies3 data frame**. For a codebook describing the variables in this dataframe, please see this website.

```
# load the data
load('movies.Rdata')

# only keep movies rated "G", "PG", "PG-13", "R"
movies3 <- movies[movies$mpaa_rating %in% c("G", "PG", "PG-13", "R"), ]
movies3$mpaa_rating <- droplevels(movies3$mpaa_rating)
```

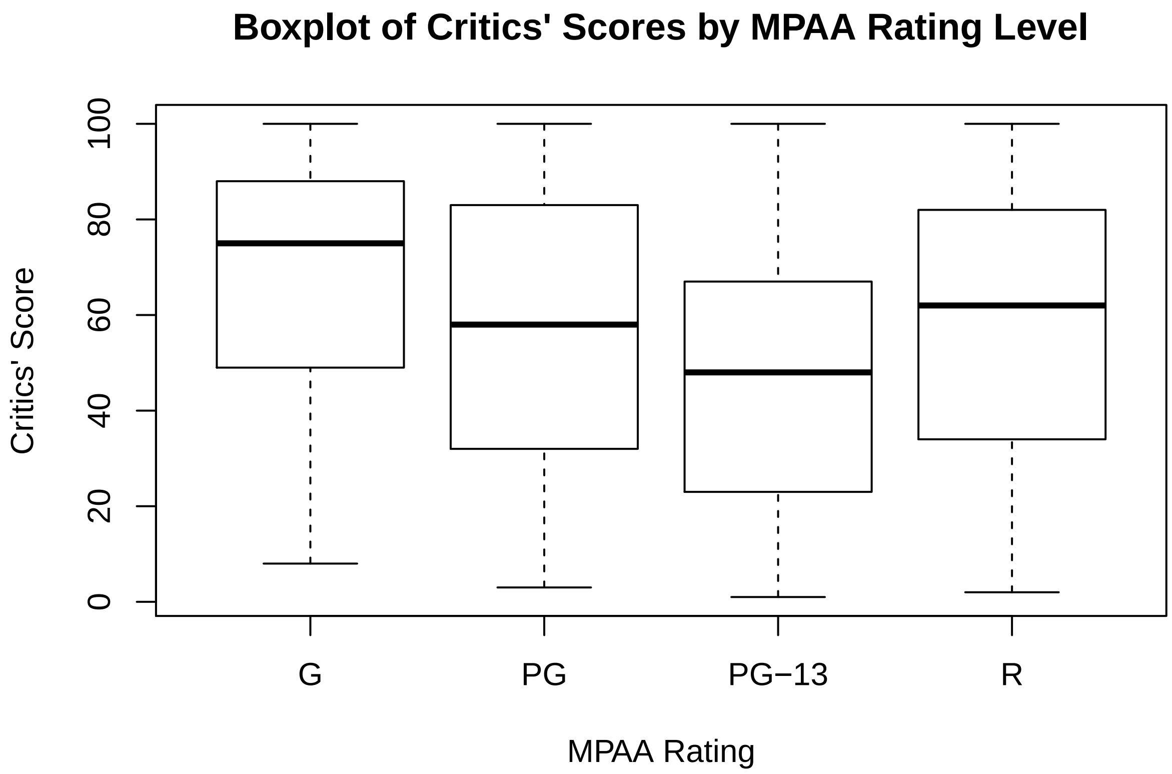
Part 1.1 (5 points) - step 0 Let's start our analysis by describing and plotting the data. Please report the number of cases and the number of variables in the `movies3` data frame, and what each case corresponds to. Also, create a side-by-side boxplot comparing the critics' scores of the movie for each MPAA rating level. Does it appear that the critics' scores differ on average depending on the MPAA classification of the movie?

```
summary(movies3)
```

```
##     title                  title_type                 genre
##  Length:599         Documentary : 23   Drama             :293
##  Class :character   Feature Film:573   Comedy            : 87
##  Mode  :character   TV Movie    :  3   Action & Adventure: 65
##                                        Mystery & Suspense: 59
##                                        Documentary       : 21
##                                        Horror            : 21
##                                        (Other)           : 53
##     runtime    mpaa_rating                              studio
##  Min.   : 40   G    : 19   Paramount Pictures              : 37
##  1st Qu.: 93   PG   :118   Warner Bros. Pictures           : 30
##  Median :103   PG-13:133   Sony Pictures Home Entertainment: 27
##  Mean   :106   R    :329   Universal Pictures              : 23
##  3rd Qu.:116               Warner Home Video               : 19
##  Max.   :202               (Other)                         :456
##                            NA's                            :  7
##  thtr_rel_year  thtr_rel_month    thtr_rel_day    dvd_rel_year
##  Min.   :1970   Min.   : 1.000   Min.   : 1.00   Min.   :1991
##  1st Qu.:1990   1st Qu.: 4.000   1st Qu.: 7.00   1st Qu.:2001
##  Median :1999   Median : 7.000   Median :15.00   Median :2003
##  Mean   :1997   Mean   : 6.783   Mean   :14.53   Mean   :2004
##  3rd Qu.:2006   3rd Qu.:10.000   3rd Qu.:21.50   3rd Qu.:2007
##  Max.   :2014   Max.   :12.000   Max.   :31.00   Max.   :2015
##                                                  NA's   :7
##  dvd_rel_month     dvd_rel_day     imdb_rating    imdb_num_votes
##  Min.   : 1.000   Min.   : 1.00   Min.   :1.900   Min.   :   390
##  1st Qu.: 3.000   1st Qu.: 7.00   1st Qu.:5.900   1st Qu.:  5576
##  Median : 6.000   Median :15.00   Median :6.500   Median : 17190
##  Mean   : 6.299   Mean   :14.96   Mean   :6.418   Mean   : 62018
##  3rd Qu.: 9.000   3rd Qu.:23.00   3rd Qu.:7.200   3rd Qu.: 64681
##  Max.   :12.000   Max.   :31.00   Max.   :9.000   Max.   :893008
##  NA's   :7        NA's   :7
##         critics_rating critics_score    audience_rating audience_score
##  Certified Fresh:118    Min.   :  1.00   Spilled:270     Min.   :11.00
##  Fresh          :178    1st Qu.: 31.00   Upright:329     1st Qu.:45.00
##  Rotten         :303    Median : 59.00                   Median :63.00
##                         Mean   : 55.41                   Mean   :61.01
##                         3rd Qu.: 80.00                   3rd Qu.:78.50
##                         Max.   :100.00                   Max.   :97.00
##
##  best_pic_nom best_pic_win best_actor_win best_actress_win best_dir_win
##  no :577      no :592      no :509        no :529          no :556
##  yes: 22      yes:  7      yes: 90        yes: 70          yes: 43
##
##
##
##
##
##  top200_box   director             actor1             actor2
##  no :584    Length:599         Length:599         Length:599
##  yes: 15    Class :character   Class :character   Class :character
##             Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     actor3             actor4             actor5
##  Length:599         Length:599         Length:599
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
```

```
##
##
##    imdb_url              rt_url
##  Length:599         Length:599
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

```
boxplot(movies3$critics_score ~ movies3$mpaa_rating,
        main = "Boxplot of Critics' Scores by MPAA Rating Level",
        xlab = "MPAA Rating", ylab = "Critics' Score")
```



Answers:

There are 599 cases (corresponding to 599 total movie titles), and 32 variables observed. It does appear that the scores from critics changes depending on the MPAA Rating: The scores for PG-13 movies seem to be lower on average, while the scores for G-rated movies seem to be higher.

Part 1.2 (5 points): Let's examine whether there is a statistically significant difference in the mean critics' scores for each MPAA level. Start by stating the null and alternative hypotheses in symbols and words, and also state the alpha level that is most commonly used.

In words

Null Hypothesis: There is no difference in the mean critics' scores for each MPAA level.

Alternative Hypothesis: There is a statistically significant difference in the mean critics' scores for each MPAA level (there is at least one difference)

In symbols

$H_0 : \mu_G = \mu_{PG} = \mu_{PG-13} = \mu_R$

$H_A : \mu_i \neq \mu_j$ for some i, j

The significance level

$\alpha = 0.05$

Part 1.3 (5 points): For our first analyses, let's use the MAD statistic that we discussed in class to compare the mean critic scores between the different MPAA rating levels. Do the following steps:

1) Extract a vector from the movies3 data frame that has the critics' scores and store it in an object called `critic_scores`
2) Extract a vector from the movies3 data frame that has the MPAA ratings and store it in an object called `MPAA_ratings`
3) Call the `get_group_means(data, grouping)` function I wrote above to get the mean of the critics' scores for each MPAA rating. Save these means in an object called `group_means`
4) Call the `get_MAD_stat(group_means)` function I wrote above to get the MAD statistic.

Report the group means values below along with the MAD statistic value.

```r
# store the critics scores and the MPAA ratings in objects
critic_scores <- movies3$critics_score
MPAA_ratings <- movies3$mpaa_rating

# get the mean critics's scores for each MPAA rating level
# Group means shown below for G, PG, PG-13, R respectively
(group_means <- get_group_means(critic_scores, MPAA_ratings))
```

```
## [1] 65.94737 56.79661 47.24812 57.59878
```

```r
# Calculate the MAD statistic
(obs_stat <- get_MAD_stat(group_means))
```

```
## [1] 9.48332
```

Part 1.4 (10 points): Now run steps 2-5 of the hypothesis test, as discussed in class. Be sure to plot the null distribution along with a red vertical line at the real MAD statistic value. Also report the p-value below. Based on this analysis comparing group means using the MAD statistic, does there appear to be a difference between between the critic's scores depending on the MPAA rating? (note: please use the answer section below to answer any questions that are posed in this and future homeworks).

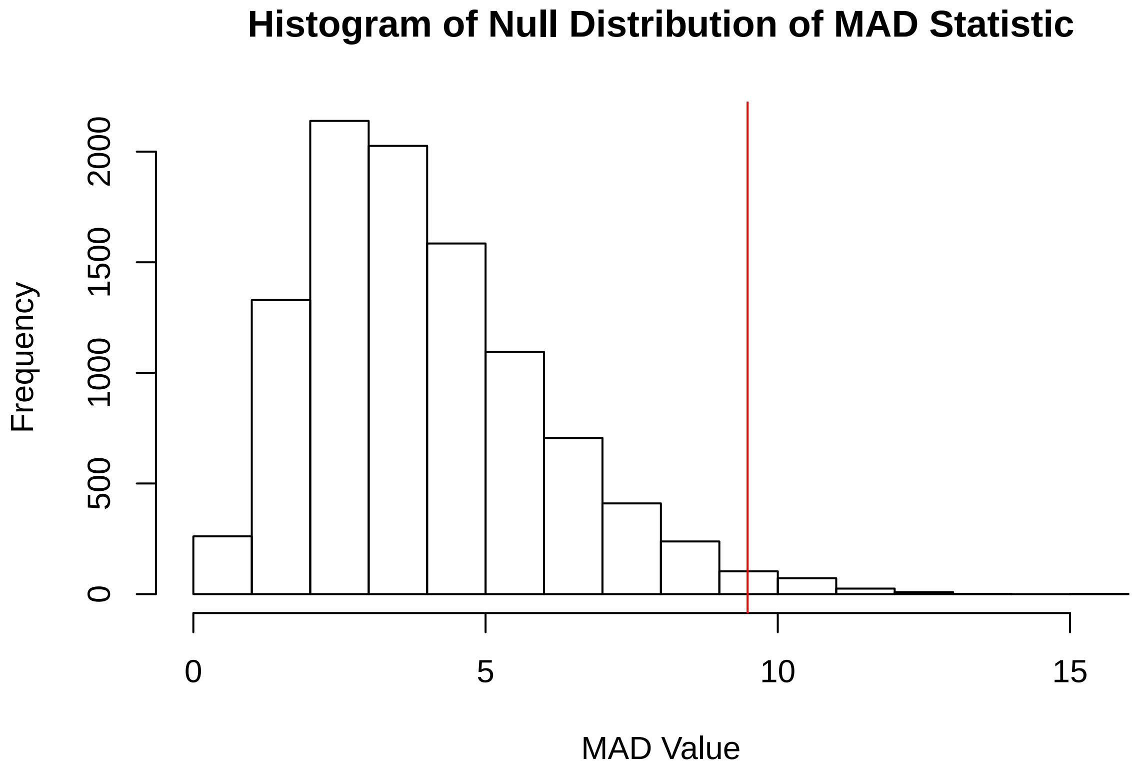
```
# create the null distribution
null_dist <- NULL
for (i in 1:10000){

  shuff_rating <- sample(movies3$mpaa_rating)
  shuff_group_means <- as.vector(by(movies3$critics_score, shuff_rating, mean))
  null_dist[i] <- get_MAD_stat(shuff_group_means)

}


# plot the null distribution with a red vertical line for the statistic value

hist(null_dist, main = "Histogram of Null Distribution of MAD Statistic",
     xlab = "MAD Value", ylab = "Frequency")
abline(v = obs_stat, col = "red")
```



```
# report the p-value

(p_val <- sum(null_dist >= obs_stat)/length(null_dist))
```

```
## [1] 0.0158
```

Answers

The p-value is 0.0158. Since this is less than our significance level $\alpha = 0.05$, there is statistically significant evidence that there is difference in the critics' scores based on MPAA rating.

Part 1.5 (10 points): We could also run the hypothesis test comparing the means based on using an F-statistic. The equation for an F-statistic is:

$$F \;=\; \frac{\text{between-group variability}}{\text{within-group variability}} \;=\; \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

In the above equation, the symbols mean the following:

- K = 4 corresponds to the 4 MPAA rating levels (G, PG, PG-13, and R)
- N corresponds to the total number of movies we are using in our analysis
- x_{ij} corresponds to the j^{th} movie with a rating in the i^{th} MPAA rating level
- n_i is the number of movies in the i^{th} group (e.g., the number of movies with a rating of G)
- $\bar{x}_i$ is the average score for the i^{th} group (e.g., the average score for movies with a rating of G)

We will discuss this equation a bit more when we discuss ANOVAs, but for now we can just think of it as a number that describes differences in the means of our data.

To calculate the F-statistic, I have written a function called `get_F_statistic(data, grouping)`. This function takes a vector of data values, and a vector indicating which group each data value belongs to.

Please rerun steps 2-5 of a hypothesis test (i.e., permutation test) using F-statistic in the R chunk below. Again be sure to plot the null distribution along with a red vertical line a the real observed F-statistic value. Also report the p-value below. Based on this analysis comparing group means using the F-statistic, does there appear to be a difference between between the critic's scores depending on the MPAA rating?

```r
# calculate the observed statistic using the get_F_statistic function
(obs_stat_F <- get_F_statistic(critic_scores, MPAA_ratings))
```

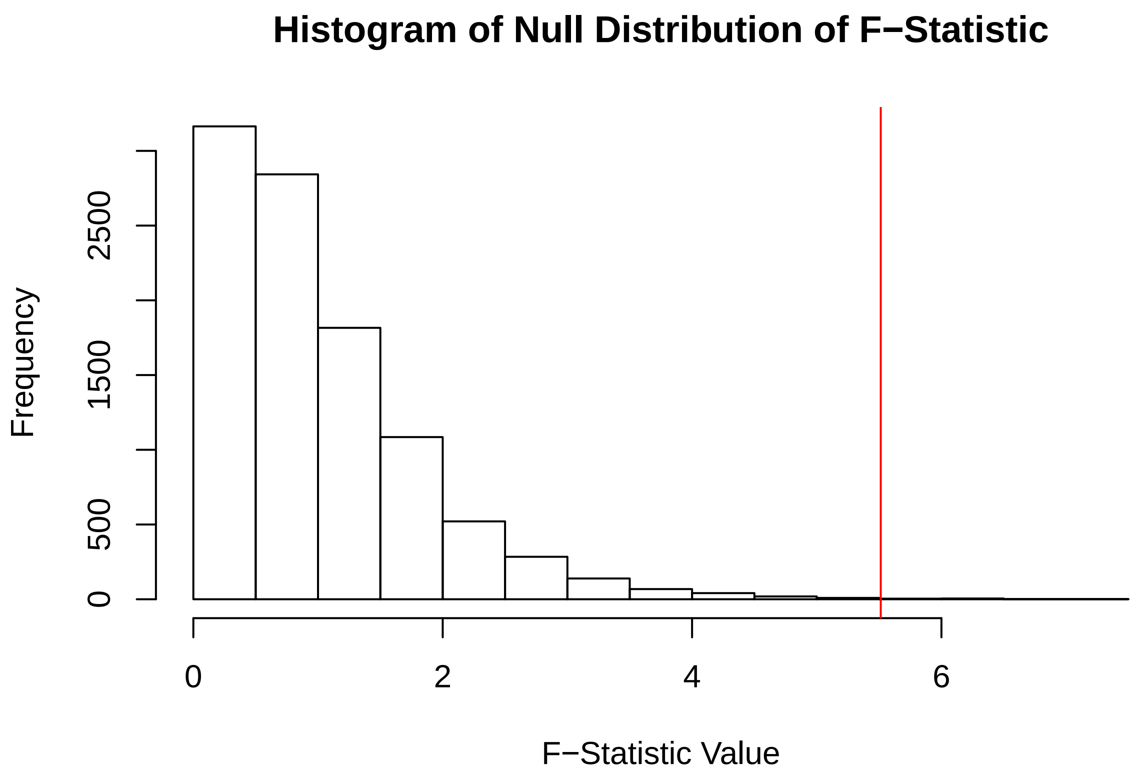
```
## [1] 5.513491
```

```r
# create the null distribution
null_dist_F <- NULL
for (i in 1:10000){

  shuff_rating <- sample(movies3$mpaa_rating)
  null_dist_F[i] <- get_F_statistic(movies3$critics_score, shuff_rating)

}


# plot the null distribution with a red vertical line for the statistic value
hist(null_dist_F, main = "Histogram of Null Distribution of F-Statistic",
     xlab = "F-Statistic Value", ylab = "Frequency")
abline(v = obs_stat_F, col = "red")
```



```r
# report the p-value

(p_val_F <- sum(null_dist_F >= obs_stat_F)/length(null_dist_F))
```

```
## [1] 0.0011
```

Answers

The p-value is 0.0011, which means we should reject our null hypothesis. There appears to be a difference in critics’ scores based on rating.

Part 1.6 (10 points): Let’s try running a permutation test using one more statistic, namely the the statistic that returns $max\ \bar{x}_i - \bar{x}_j$, where $\bar{x}_i$, and $\bar{x}_j$ refer to mean critic score for the i^{th} and j^{th} movie rating levels. To do this analysis, start by writing a function yourself called ’get_max_diff(data, grouping)‘ that takes a data vector (i.e., the critics ratings) and a grouping vector (i.e., the MPAA ratings), and returns the maximum value for the difference between the means over all pairs of groups.

Hints: One way you can write this by modifying the `get_MAD_stat` function (e.g., start by copying and pasting it to the chunk below). Again use two nested for loops, but instead of summing the results, use an if statement and to store the difference between mean scores if the current difference is greater than any difference seen on previous iterations (i.e., create an object called `max_diff` that initially has a value of 0, and update this value when in any interation of the loop that has a greater value for the difference between means).

Once you have written this function, calculate and report the observed statistic value using this function.

```
# a function to the maximum absolute difference between all pairs of means
get_max_diff <- function(data, grouping) {
  max_diff <- 0
  means_vec <- as.vector(by(data, grouping, mean))

  for (iGroup1 in 1:(length(means_vec) - 1)) {

    for (iGroup2 in (iGroup1 + 1):(length(means_vec))){
      diff <- abs(means_vec[iGroup1] - means_vec[iGroup2])
      if(diff > max_diff){
        max_diff <- diff
      }
    }
  }
    max_diff
}

# end of the function


# use the function to get the observed statistic

(obs_stat_max <- get_max_diff(critic_scores, MPAA_ratings))
```

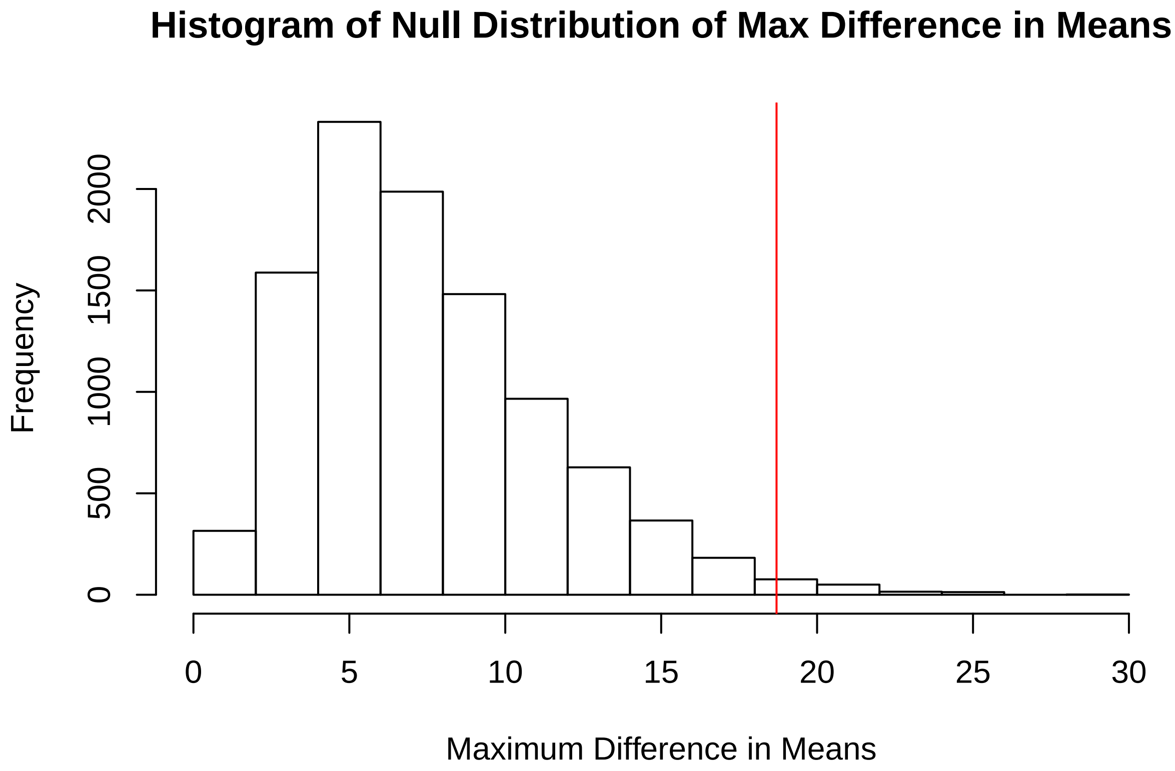
```
## [1] 18.69925
```

Part 1.7 (10 points): Now Repeat steps 2-5 of hypothesis testing using 'get_max_diff(data, grouping)` function. Again be sure to plot the null distribution along with a red vertical line a the real observed max-diff statistic value, report the p-value below, and answer the question about whether based on this analysis comparing group means using the max-diff statistic, does there appear to be a difference between between the critic's scores depending on the MPAA rating?

```
# create the null distribution
null_dist_max <- NULL
for (i in 1:10000){
  shuff_ratings <- sample(movies3$mpaa_rating)
  null_dist_max[i] <- get_max_diff(movies3$critics_score, shuff_ratings)
}

# plot the null distribution with a red vertical line for the statistic value

hist(null_dist_max, main = "Histogram of Null Distribution of Max Difference in Means",
     xlab = "Maximum Difference in Means", ylab = "Frequency" )
abline(v = obs_stat_max, col = "red")
```



```r
# report the p-value
(p_val_max <- sum(null_dist_max >= obs_stat_max)/length(null_dist_max))
```

```
## [1] 0.0118
```

Answers

The p-value is 0.0118, which is less than significance level $\alpha = 0.05$. Based on this analysis there does seem to be a difference in critic scores based on MPAA rating.

Part 1.8 (5 points): In the three exercises above, you ran three permutation tests based on three different statsitics. Describe which statistic/permutation test seems best and why?

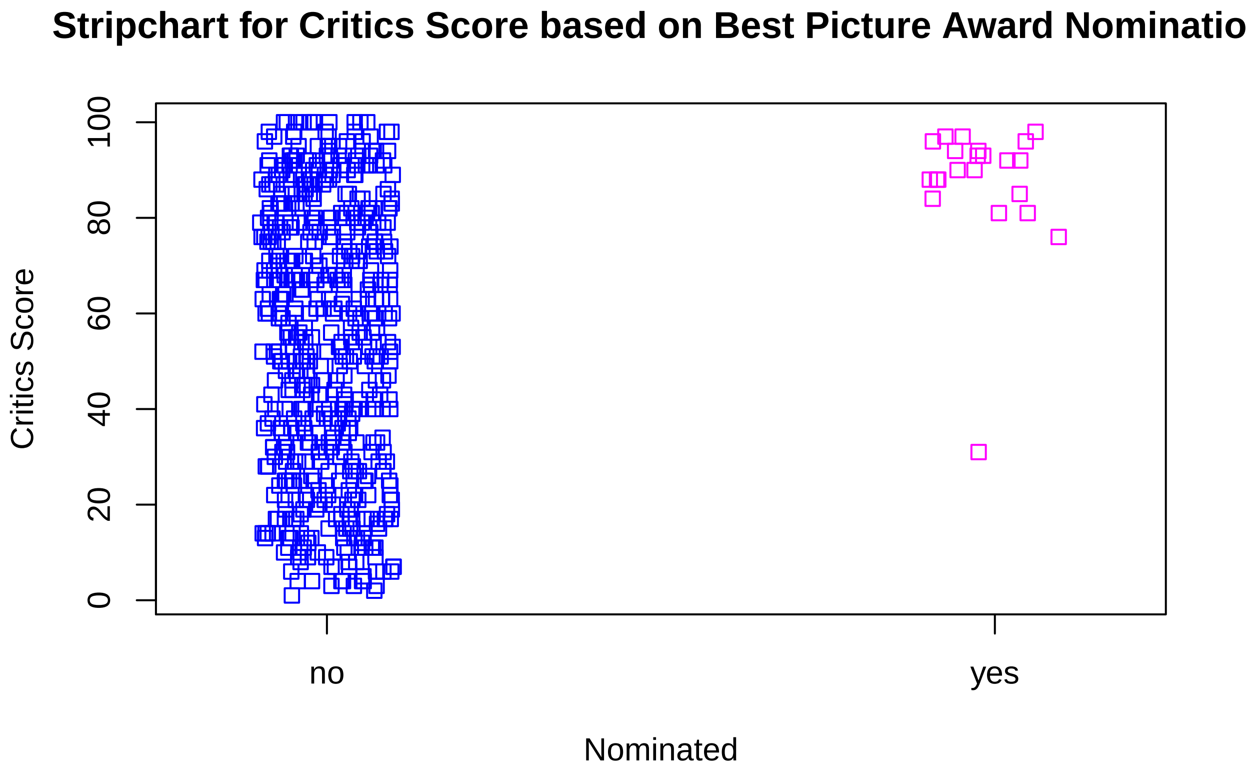
Answers

From these three permutation tests it seems that the F-statistic test is best. With the maximum difference test, it seems that the variance in means could easily be overestimated by outlier statistics, since we are only looking at the difference between the means of two groups instead of all of them. With the Mean Absolute Difference, this difference is slightly diminished because we are considering the difference between the means of each group. However, the F-statistic seems to be the most comprehensive because it takes into account both the variability between the group means and variability within the group. From the p-values we can also see that the F-statistic test yielded the strongest result (lowest p-value), which also suggests strength as a test.

Part 1.9 (5 points): Using the R chunk below, explore the movies data more and create one additional plot that shows something interesting.

```
#tbl_df(movies3) #just to visualize a table of the movies data, not shown

#I created a stripchart comparing the scores from critics
#based on whether they received a "Best Picture" nomination
stripchart(movies3$critics_score ~ movies3$best_pic_nom,
           vertical = T, col = c("blue", "magenta"), method = "jitter",
           main = "Stripchart for Critics Score based on Best Picture Award Nomination",
           xlab = "Nominated",
           ylab = "Critics Score")
```



Problem 2: Permutation tests for correlation

Part 1: The 1969 draft lottery

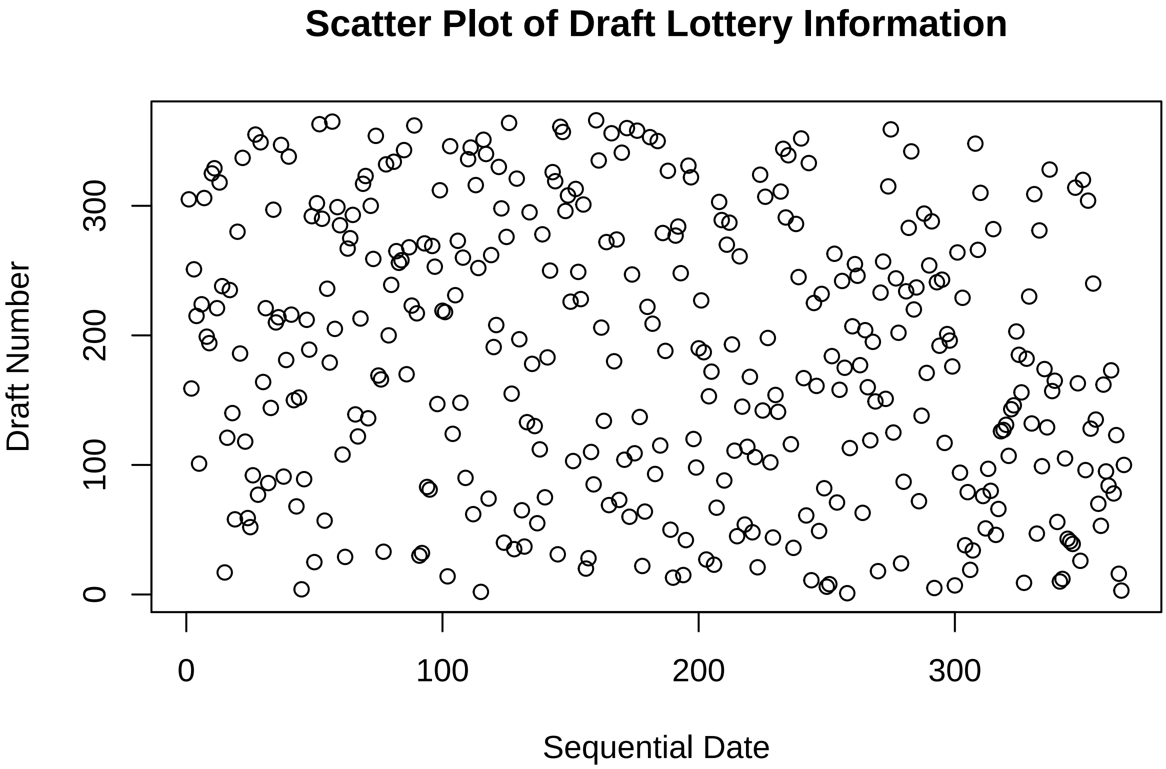
In 1969, the United States Selective Service conducted a lottery to decide which young men would be drafted into the armed forces. Each of the 366 birthdays in a year (including February 29) was assigned a draft number. Young men born on days that were assigned low draft numbers were drafted.

If the draft was completely fair, there should be no correlation between the draft number and the date someone was born. In the following set of exercises we will use hypothesis testing to assess whether there was indeed no correlation.

Part 2.0 (2 points): Let's start our analyses, as usual, by visualizing the data. A data frame that contains the draft lottery information can be loaded into R using the code below. This data frame contains two variables (columns). The first column contains sequential days of the year and the second column contains the draft number associated with that date. Create a scatter plot of the draft number as a function of the sequential date. Does there appear to be any trend in the data?

```r
# load the data into R
load('draft_lottery_data.Rda')

# plot the data
plot(draft_lottery_data$Draft_Number ~ draft_lottery_data$Sequential_Date,
     main = "Scatter Plot of Draft Lottery Information",
     xlab = "Sequential Date",
     ylab = "Draft Number")
```



Answer:

Looking at the scatter plot there does not appear to be any trend in the data.

Part 2.1 (3 points): Now let's do step 1 of our null hypothesis significance tests (NHSTs) by stating the null and alternative hypotheses in symbols and in words, and state the significance level.

Answer:

In Words:

Null hypothesis: There is no correlation between the date picked and the draft number assigned.

Alternative hypothesis: There is a correlation between the date picked and the draft number assigned.

In Symbols:

$\mu_0 : \rho = 0$

$\mu_A : \rho \neq 0$

Part 2.2 (2 points): Next let's do step 2 of hypothesis testing by calculating the statistic of interest and save it to a variable obs_stat. Describe what this statistic means as clearly as you can (e.g., if the statistic is negative what does that mean in terms of dates and draft numbers?).

```r
(obs_stat <- cor(draft_lottery_data$Draft_Number,
                 draft_lottery_data$Sequential_Date))
```

```
## [1] -0.2260414
```

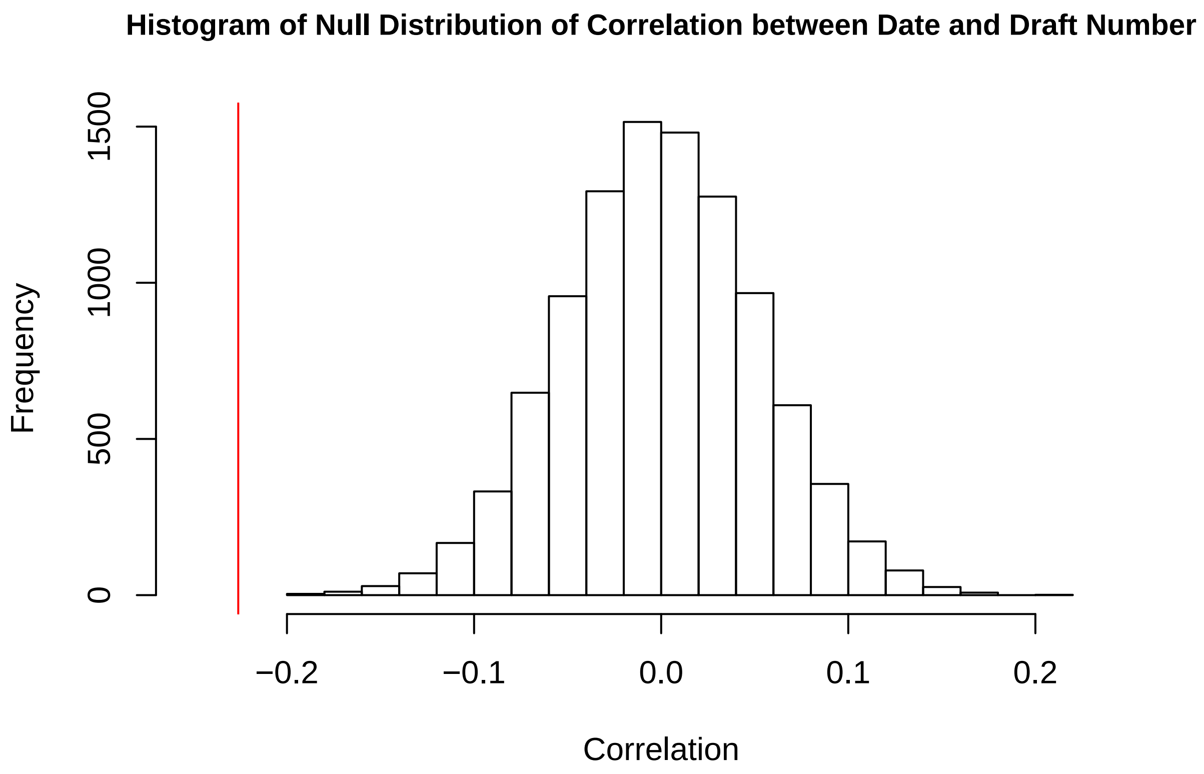
Answer:

The observed correlation is -0.2260414. This means there is a moderate negative correlation between Draft Number and Date of Birth, so the later the sequential date, the lower the draft number was likely to be.

Part 2.3 (5 points): Now let's do step 3 of hypothesis testing by creating a null distribution. You can calculate one point in the null distribution by shuffling one of the variables (using the `sample()` function) and then calculating the correlation. Use a for loop to repeat this process 10,000 times to generate the full null distribution. Plot a histogram of the null distribution, put a red veritical line on it at the value of the observed statistic, and describe the null distributions shape. Is the center of the null distribution at a value that makes sense to you?

```r
# create the null distribution and plot it
null_dist <- NULL
for (i in 1:10000){
  shuff_data <- sample(draft_lottery_data$Sequential_Date)
  null_dist[i] <- cor(draft_lottery_data$Draft_Number, shuff_data)
}

hist(null_dist,
     main = "Histogram of Null Distribution of Correlation between Date and Draft Number",
     xlab = "Correlation", ylab = "Frequency", xlim = c(-0.25, 0.25), cex.main = 0.9)
abline (v = obs_stat, col = "red")
```



Answer:

The shape of the null distribution is approximately normal. The center is at approximately 0, which makes sense because if there is no correlation then $\rho = 0$. This is acheived by shuffling (randomizing) the data to be used in calculating correlation, so there should be the most calculated correlations close to 0.

Part 2.4 (5 points): Now use the vector null_dist and the obs_stat to calculate the p-value by seeing the proportion of points in the null distribution that are *as extreme or more extreme* than the observed statistic. Is this p-value consistent with there being no correlation between draft numbers and sequential dates?

```
(p_val <- sum(abs(null_dist) >= abs(obs_stat))/length(null_dist))
```

```
## [1] 0
```

Answer:

The p-value is equal to 0, thus we should reject the null hypothesis. There is evidence that there is some correlation between date and draft number.

Part 2.5 (5 points): Make a judgement call as to whether you believe the draft lottery was fair. Make sure to justify your answer.

Answer:

I don't believe the draft lottery was fair because according to our analysis, the probability of such results occuring (that the correlation is -0.22) is zero. Since the numbers were physically drawn from slips of paper (Source: Wikepedia) it is possible that the papers were put into a box in a certain order and not mixed well enough to acheive a random distribution.

Part 2.6 (10 points): Calculate a 95% confidence interval for the value of the correlation between sequential date and draft number using the bootstrap. Note that you can sample points in a *data frame* with replacement using: bootstrap_data_frame <- sample(draft_lottery_data, size = 366, replace = TRUE). Does the confidence interval contain 0, and would you expect it to contain 0?

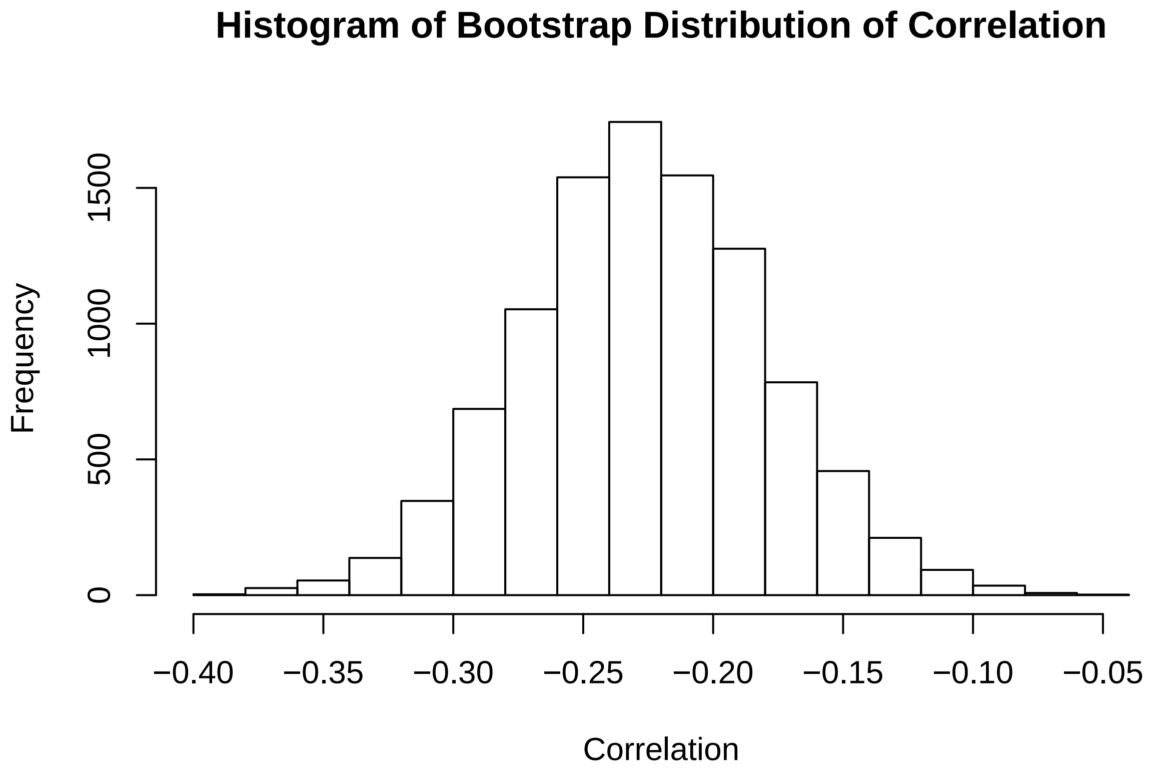
```
 # an example of a bootstrapped data frame
 one_bootstrap_data_frame <- draft_lottery_data[sample(1:366, 366, replace = TRUE), ]



 # Use a for loop to create a full bootstrap distribution
bootstrap_dist <- NULL
for (i in 1:10000){
 one_bootstrap_data_frame <- draft_lottery_data[sample(1:366, 366, replace = TRUE), ]
  bootstrap_dist[i] <- cor(one_bootstrap_data_frame$Sequential_Date,
                           one_bootstrap_data_frame$Draft_Number)
}



 # plot the bootstrap distribution
 hist(bootstrap_dist, main = "Histogram of Bootstrap Distribution of Correlation",
      xlab = "Correlation", ylab = "Frequency")
```



```
# create confidence intervals based on the bootstrap
(CI <- obs_stat + sd(bootstrap_dist)*c(-2,2))
```

```
## [1] -0.3200475 -0.1320353
```

Answer:

The confidence interval here is [-0.3200475, -0.1320353]. It does not contain zero, and we do not expect it to since the confidence interval should contain the true population parameter 95% of the time, and we know it is extremely unlikely that 0 is the population parameter (this would imply no correlation between Date and Draft Number).

Reflection (3 points)

Please reflect on how the homework went by going to Canvas, going to the Quizzes link, and clicking on Reflection on homework 5